

SSA-based approaches to analysis and forecast of multidimensional time series

N. Golyandina, D. Stepanov

St. Petersburg University, Mathematical Department

Abstract

Reconstruction and forecast of multidimensional signals in the presence of noise is considered. SSA-based methods (MSSA, CSSA) are applied. Comparison of approaches and investigation of their features are performed by means of statistical simulation.

Introduction

Let us consider the problem of signal reconstruction and forecasting for a system of simultaneous time series. The Singular Spectrum Analysis (SSA) method [1, 2] and its extensions for processing multidimensional time series (Multi-channel SSA (MSSA) and Complex SSA (CSSA) [3, 4]) can be used for solving this problem.

The questions concerning choice of parameters for signal reconstruction and forecasting, relation between optimal parameters for analysis and forecast, comparison of accuracies of different SSA-based methods arise. Unfortunately, only few of these questions have theoretical answer at the moment due to complicated nonlinear construction of methods. That is why we use statistical simulation here. Namely, we simulate time series consisting of a signal and of white noise, then extract signal, forecast it and finally study the estimated errors. It is important to choose a proper class of considered signals to obtain meaningful results.

The first step of SSA-based methods is the embedding procedure: transformation of the time series to a sequence of so-called lagged vectors. The linear space spanned by these lagged vectors is called the *trajectory space*. Just the trajectory space (its dimension, form of basis vectors) determines the time series structure from the viewpoint of SSA. Generally, trajectory spaces can vary for different multivariate extensions of SSA. In some sense the more complex time series structure leads to the bigger dimension of the trajectory space.

Consider a system $H^{(k)} = (h_j^{(k)})_{j=0}^{N-1}$, $k = 1, \dots, s$, of s signals with lengths N . Let r_k denote the trajectory dimension of $H^{(k)}$ (i.e., dimension of the trajectory spaces generated by onedimensional SSA applied to this time series) and r denote the multivariate trajectory dimension generated by MSSA applied to the time series system as a whole. Relation between r and r_k , $k = 1, \dots, s$, is studied in [4]. In particular, it is shown that $r_{\min} \leq r \leq r_{\max}$, where $r_{\min} = \max\{r_k, k = 1, \dots, s\}$

and $r_{\max} = \sum_{k=1}^s r_k$. The case $r = r_{\max}$ is worst for MSSA and means that the time series don't contain matched components. The case $r < r_{\max}$ indicates the presence of matched components and can lead to advantages of simultaneous processing of the time series system.

Section 1 contains a brief description of the used SSA-based algorithms (see the full description of algorithms and the developed theory in [2]–[4]). Section 2 provides results of simulation applied to systems of two series of noised harmonic signals. Since the complex-valued CSSA method can be applied to a system of only two real-valued time series, we consider the case $s = 2$ to include CSSA into the scope of the paper.

1 Algorithms

Consider a system $F^{(k)} = (f_j^{(k)})_{j=0}^{N-1}$, $k = 1, \dots, s$, of s time series with length N . This section contains a brief description of SSA-based algorithms used for extraction (reconstruction) of the signals and their forecasting.

The algorithms are formulated only for MSSA since SSA is its particular case for $s = 1$ and CSSA is a natural transfer of SSA to the complex-valued case.

1.1 MSSA analysis

1st step: Embedding

Let L be an integer (*window length*), $1 < L < N$. For each time series $F^{(k)}$ the embedding procedure forms $K = N - L + 1$ lagged vectors $X_j^{(k)} = (f_{j-1}^{(k)}, \dots, f_{j+L-2}^{(k)})^T$, $1 \leq j \leq K$. The *trajectory matrix* of the multidimensional series $(F^{(1)}, \dots, F^{(s)})$ is a matrix $L \times Ks$ and has the form

$$\mathbf{X} = [X_1^{(1)} : \dots : X_K^{(1)} : \dots : X_1^{(s)} : \dots : X_K^{(s)}] = [\mathbf{X}^{(1)} : \dots : \mathbf{X}^{(s)}].$$

The *trajectory space* is a linear space spanned by lagged vectors (columns of the trajectory matrix).

2nd step: Singular Value Decomposition (SVD)

Let $\mathbf{S} = \mathbf{X}\mathbf{X}^T$, $\lambda_1 \geq \dots \geq \lambda_L \geq 0$ be *eigenvalues* of the matrix \mathbf{S} , $d = \max\{j : \lambda_j > 0\}$, U_1, \dots, U_d be the corresponding *eigenvectors*, and $V_j = \mathbf{X}^T U_j / \sqrt{\lambda_j}$, $j = 1, \dots, d$, be *factor vectors*. Denote $\mathbf{X}_j = \sqrt{\lambda_j} U_j V_j^T$. Then the SVD of the trajectory matrix \mathbf{X} can be written as

$$\mathbf{X} = \mathbf{X}_1 + \dots + \mathbf{X}_d. \tag{1}$$

3rd step. Grouping

Once the expansion (1) has been obtained, the grouping procedure partitions the set of indices $\{1, \dots, d\}$ into m disjoint subsets I_1, \dots, I_m . Let $I = \{i_1, \dots, i_p\}$. Then the *resultant matrix* \mathbf{X}_I corresponding to the group I is defined as $\mathbf{X}_I = \mathbf{X}_{i_1} + \dots + \mathbf{X}_{i_p}$. Thus, we have the grouped decomposition

$$\mathbf{X} = \mathbf{X}_{I_1} + \dots + \mathbf{X}_{I_m}. \tag{2}$$

4th step: Diagonal averaging

The last step is in a sense opposite to the first step and transforms each matrix of the grouped decomposition (2) into a system of new (reconstructed) series of length N by hankelization-like procedure (see the formal description in [2, 3]).

Thus, the result of SSA-based algorithms is an expansion of the (multidimensional) time series to sum of m series; parameters are the window length L and the way of grouping. Note also that the case of two series components ($m = 2$) is often more sensibly regarded as the problem of separating out a single component (for example, as a noise reduction) rather than the problem of separation of two terms. In this case, we are interested in only one group of indices related to the signal, namely I_1 . Designate by $\tilde{F}^{(k)} = (\tilde{f}_j^{(k)})_{j=0}^{N-1}$, $k = 1, \dots, s$ the reconstructed time series corresponding to the signal group of indices.

1.2 MSSA forecast

Let the leading r eigentriples (λ_j, U_j, V_j) be identified and chosen as related to the signal (r is treated as the signal trajectory dimension and $I_1 = \{1, \dots, r\}$). Then the algorithm of MSSA analysis gives us the system of s reconstructed signals $\tilde{F}^{(k)} = (\tilde{f}_j^{(k)})_{j=0}^{N-1}$, $k = 1, \dots, s$. Denote by $R_N = (\tilde{f}_N^{(1)}, \tilde{f}_N^{(2)}, \dots, \tilde{f}_N^{(s)})^T$ the vector of forecasted signal values for each time series from the system. Below we rewrite forecasting formulae for two variants of MSSA forecast: MSSA- L (generated by $\{U_j\}_{j=1}^r$) and MSSA- K (generated by $\{V_j\}_{j=1}^r$). These one-term ahead forecasting formulae can be applied to M -term ahead forecast by recurrence. Note that the SSA forecast coincides with the MSSA- L forecast for $s = 1$.

1.3 MSSA- L

Denote by \mathbf{Y} the matrix consisting of the last $L - 1$ values of the reconstructed signals:

$$\mathbf{Y} = \begin{pmatrix} \tilde{f}_{N-L+1}^{(1)}, \dots, \tilde{f}_{N-1}^{(1)} \\ \tilde{f}_{N-L+1}^{(2)}, \dots, \tilde{f}_{N-1}^{(2)} \\ \vdots \\ \tilde{f}_{N-L+1}^{(s)}, \dots, \tilde{f}_{N-1}^{(s)} \end{pmatrix},$$

by U_j^∇ the vectors of the first $L - 1$ coordinates of the eigenvectors U_j , by π_j the last coordinates of the eigenvectors and, finally, $\nu = \sum_{j=1}^r \pi_j^2$. If $\nu < 1$, then the MSSA- L forecast exists and can be calculated by the formula

$$R_N = \mathbf{Y} \mathcal{R}_L, \quad \text{where} \quad \mathcal{R}_L = \frac{1}{1 - \nu^2} \sum_{j=1}^r \pi_j U_j^\nabla \in \mathbb{R}^{L-1}. \quad (3)$$

Note that the formula (3) means forecasting each of signals by the same linear recurrent formula generated by the whole system.

1.4 MSSA-K

Let $V_j^{\nabla s} \in \mathbb{R}^{(K-1)s}$ be the vectors consisting of all coordinates of the V_j but coordinates with numbers Kp (we denote them $\pi_j^{(p)}$), $p = 1, \dots, s$. Introduce vectors of the last $K - 1$ values of the reconstructed signals

$$\mathcal{Z}^{(m)} = (\tilde{f}_{N-K+1}^{(m)}, \dots, \tilde{f}_{N-1}^{(m)})^T, \quad m = 1, \dots, s,$$

and denote $\mathbf{Q} = (V_1^{\nabla s}, V_2^{\nabla s}, \dots, V_r^{\nabla s})$,

$$\mathcal{Z} = \begin{pmatrix} \mathcal{Z}^{(1)} \\ \mathcal{Z}^{(2)} \\ \vdots \\ \mathcal{Z}^{(s)} \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} \pi_1^{(1)} & \pi_2^{(1)} & \dots & \pi_r^{(1)} \\ \pi_1^{(2)} & \pi_2^{(2)} & \dots & \pi_r^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_1^{(s)} & \pi_2^{(s)} & \dots & \pi_r^{(s)} \end{pmatrix}.$$

If the inverse matrix $(\mathbf{I}_{ss} - \mathbf{W}\mathbf{W}^T)^{-1}$ exists and $r \leq (K - 1)s$, then the MSSA-K forecast exists and can be calculated by the formula

$$R_N = (\mathbf{I}_{ss} - \mathbf{W}\mathbf{W}^T)^{-1} \mathbf{W}\mathbf{Q}^T \mathcal{Z}. \tag{4}$$

Note that the formula (4) means forecasting signals by some multidimensional linear recurrent formula.

2 Numerical investigation

Let us observe $(F^{(1)}, F^{(2)}) = (H^{(1)}, H^{(2)}) + (N^{(1)}, N^{(2)})$, where $(H^{(1)}, H^{(2)})$ is a two-dimensional signal consisting of two harmonic time series, $N^{(1)}$ and $N^{(2)}$ are realizations of independent normal white noises. Then we can use standard simulation procedure to obtain estimates of mean square errors (MSE) for reconstruction and forecasting of $(H^{(1)}, H^{(2)})$ by the considered above SSA-based methods. Note that the resultant MSE is calculated as sum of $\text{MSE}^{(1)}$ and $\text{MSE}^{(2)}$ for $H^{(1)}$ and $H^{(2)}$ correspondingly.

We take the following parameters for generation of time series: $N = 71$, variance of noises $\sigma^2 = 25$. Number of realization is equal to 10000.

We consider three variants of the signals $(H^{(1)}, H^{(2)})$:

Example 1 (the same periods, difference between phases not equal to $\pi/2$):

$$h_k^{(1)} = 30 \cos(2\pi k/12), \quad h_k^{(2)} = 20 \cos(2\pi k/12 + \pi/4), \quad k = 0, \dots, N - 1.$$

Example 2 (the same periods and amplitudes; phases difference equal to $\pi/2$):

$$h_k^{(1)} = 30 \cos(2\pi k/12), \quad h_k^{(2)} = 30 \cos(2\pi k/12 + \pi/2), \quad k = 0, \dots, N - 1.$$

Example 3 (different periods):

$$h_k^{(1)} = 30 \cos(2\pi k/12), \quad h_k^{(2)} = 20 \cos(2\pi k/8 + \pi/4), \quad k = 0, \dots, N - 1.$$

Choice of these examples is determined by different dimensions of signal trajectory spaces for different variants of SSA-based methods (see Table 1).

Table 1: Dimensions of signal trajectory spaces

	Example 1	Example 2	Example 3
SSA	2	2	2
MSSA	2	2	4
CSSA	2	1	4

Results of investigation for different window lengths L are summarized in Tables 2 and 3. Minimum values in rows are marked by bold font. The 24 term-ahead forecast was performed. We omit results of the CSSA forecast and Example 2 for brevity. Comparison of Tables 2 and 3 with Table 1 clearly demonstrates relation

Table 2: MSE of signal reconstruction

Example 1	$L = 12$	$L = 24$	$L = 36$	$L = 48$	$L = 60$
SSA	6.58	4.09	4.06	4.09	6.58
MSSA	6.44	3.71	3.22	3.01	4.06
CSSA	6.58	4.09	4.07	4.09	6.58
Example 2	$L = 12$	$L = 24$	$L = 36$	$L = 48$	$L = 60$
SSA	6.57	4.08	4.05	4.08	6.57
MSSA	6.44	3.71	3.24	3.03	4.02
CSSA	3.22	2.08	2.07	2.08	3.22
Example 3	$L = 12$	$L = 24$	$L = 36$	$L = 48$	$L = 60$
SSA	6.42	3.99	3.96	3.99	6.42
MSSA	13.77	7.57	6.13	5.75	7.66
CSSA	13.91	8.16	7.68	8.16	13.91

Table 3: MSE of signal forecast

Example 1	$L = 12$	$L = 24$	$L = 36$	$L = 48$	$L = 60$
MSSA-L	10.69	7.12	7.26	7.35	8.61
MSSA-K	12.00	8.26	7.46	6.57	7.80
SSA	14.32	11.00	12.26	12.62	15.65
Example 3	$L = 12$	$L = 24$	$L = 36$	$L = 48$	$L = 60$
MSSA-L	50.58	14.58	15.19	14.69	17.96
MSSA-K	39.22	16.91	16.00	13.22	16.30
SSA	14.59	11.11	12.65	12.77	16.01

between accuracy of signal reconstruction/forecast and dimension of the signal trajectory space.

Table 2 contains MSE averaging by time series points. Fig. 1 and 2 demonstrate dependence of error (of mean square deviation (MSD) equal to square root of MSE) on point number for Example 2. Fig. 1 shows that advantage of the window length $L = 48$ for MSSA in comparison with $L = 36$ is achieved due to better description

of points far from the middle. Fig. 2 demonstrates different spreading of errors for different methods and the same window length $L = 36$.

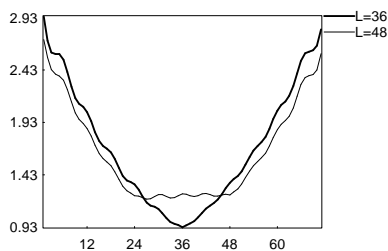


Figure 1: $MSD^{(1)}$ for different L

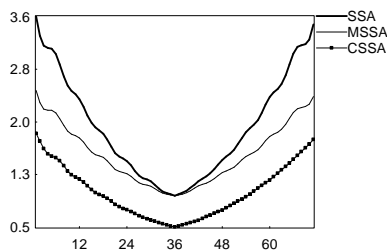


Figure 2: $MSD^{(1)}$ for different methods

Conclusions.

1. Accuracy of SSA-based methods is closely related to structure of the signal trajectory spaces generated by these methods.
2. The MSSA method has an advantage if the time series include matched components.
3. Optimal window lengths for analysis and forecast can differ.
4. Accuracy of forecast globally corresponds to accuracy of reconstruction; however the previous item shows that this relation isn't unambiguous.

References

- [1] J. Elsner, A. Tsonis (1996) *Singular Spectrum Analysis. A New Tool in Time Series Analysis*. New York: Plenum Press, 163 P.
- [2] N.Golyandina, V.Nekrutkin, A.Zhigljavsky (2001) *Analysis of Time Series Structure: SSA and Related Techniques*, Chapman & Hall/CRC, 305 P.
- [3] N.Golyandina, V.Nekrutkin, D.Stepanov (2003) Variants of the “Caterpillar”-SSA method for analysis of multidimensional time series. In: Proceedings of the II International conference “System identification and control problems” SICPRO’03, Moscow, p. 2139-2168 (in Russian).
- [4] D.Stepanov, N.Golyandina (2005) Variants of the “Caterpillar”-SSA method for forecasting multidimensional time series. In: Proceedings of the IV International conference “System identification and control problems” SICPRO’05, Moscow, p. 1831-1848 (in Russian).